On the Size Distribution of Autonomous Systems

Marwan Fayed, Paul Krapivsky, John Byers, Mark Crovella,
David Finkel, Sid Redner*
Computer Science Department
Boston University
BUCS-TR-2003-001

January 17, 2003

Abstract

This paper explores reasons for the high degree of variability in the sizes of ASes that have recently been observed, and the processes by which this variable distribution develops. AS size distribution is important for a number of reasons. First, when modeling network topologies, an AS size distribution assists in labeling routers with an associated AS. Second, AS size has been found to be positively correlated with the degree of the AS (number of peering links), so understanding the distribution of AS sizes has implications for AS connectivity properties. Our model accounts for AS births, growth, and mergers. We analyze two models: one incorporates only the growth of hosts and ASes, and a second extends that model to include mergers of ASes. We show analytically that, given reasonable assumptions about the nature of mergers, the resulting size distribution exhibits a power law tail with the exponent independent of the details of the merging process. We estimate parameters of the models from measurements obtained from Internet registries and from BGP tables. We then compare the models solutions to empirical AS size distribution taken from Mercator and Skitter datasets, and find that the simple growth-based model yields general agreement with empirical data. Our analysis of the model in which mergers occur in a manner independent of the size of the merging ASes suggests that more detailed analysis of merger processes is needed.

Keywords: autonomous systems, size distribution, AS degree, growth, merger, measurement & inference.

*M. Fayed, J. Byers and M. Crovella are with the Dept. of Computer Science at Boston University and are supported in part by NSF grant ANIR-9986397 and NSF CAREER award ANIR-0093296. E-mail: {mfayed,byers,crovella}@cs.bu.edu. P. L. Krapivsky and S. Redner are with the Dept. of Physics, as well as the Center for BioDynamics and the Center for Polymer Studies, at Boston University and are supported by grants NSF DMR9978902 and ARO DAAD19-99-1-0173. E-mail: {paulk,redner}@bu.edu. D. Finkel is with the Dept. of Computer Science at Worcester Polytechnic Institute. E-mail: dfinkel@cs.wpi.edu.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
17 JAN 2003		**17-01-2003 to 17-01-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
On the Size Distribution of Autonomous Systems	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Research Office,PO Box 12211,Research Triangle Park,NC,27709-2211**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**16**	

1 Introduction

One useful abstraction for the structure of Internet is a labelled graph. In this framework, graph nodes are routers and end systems, edges are network links, and a node's label corresponds to the autonomous system (AS) to which the node belongs. This abstraction provides an attractive target for Internet topology generation efforts, because it makes possible significant improvements in network simulation. An AS-labelled graph would allow meaningful simulation of traffic flow patterns, which are influenced strongly by interdomain routing policies. Furthermore, accurate AS labelling would allow realistic simulations of the BGP system, which is of considerable current interest.

Unfortunately, a number of gaps in our understanding prevent the construction of such a labelled graph. Principal among these is the current lack of existence of a model for the evolution of the interdomain system. Such a model should be able to answer the questions: By what processes do new ASes arise over time? How do ASes grow? By what processes do ASes merge? In this paper we seek answers to these questions. To do so we use a variety of Internet measurements, supplemented with insight from analytic models.

We start with two observations about growth in the Internet: over the past 10 years or so, the network appears to have been growing exponentially both in terms of the number of hosts in the network, and in terms of the number of autonomous systems present. Furthermore, we argue the average size of an autonomous system has been growing with time, which suggests that any realistic model *cannot* be a scale-free system (in the sense of the Barabási-Albert model for node degree [1]).

We first posit the simplest model that incorporates all of these characteristics, and explore its properties. We show that in the asymptotic time limit, this model leads to a stationary AS size distribution. In fact, we obtain the entire size distribution in the form of a recurrence relation, and we further analyze how the growth parameters determine the tail exponent.

The principal validation of our model is to check whether its predicted size distribution agrees with empirical measurements of AS size distributions. For this purpose we use two large router inventories, from the Mercator and Skitter projects, and map each router to its associated AS. The resulting size distributions of ASes are found to have long (though not clearly power-law) tails. Agreement between the AS size distributions predicted by the growth model and the data is a good first order approximation, but there are noticeable discrepancies. For example, the tail of the model distribution is in general agreement with data, but it strictly follows a power-law, while the empirical data shows some deviation from a power-law tail.

We hypothesize that an important factor our simple model omits is the merger of ASes. Statistical mechanics shows that highly variable size distributions can also result from coalescence processes (as in the formation of raindrops or polymer aggregates [6]). To understand the merger process and estimate a corresponding rate, we develop a heuristic and apply it to examinations of BGP table collections over two one-year periods. Adding mergers to our growth model complicates the analysis considerably. Currently we

have solved only the most tractable version of this model, in which mergers occur with a rate proportional to the number of ASes present, in a manner independent of the sizes of the ASes merging. This initial model exhibits improved agreement with data with respect to small to medium sized ASes, but predicts large AS sizes less well, compared to the pure growth model. More importantly, it points to the need for analysis of more realistic merger processes (such as those that account for the relative sizes of the ASes being merged).

Thus, this paper shows that a growth based model is a good first step to understanding the evolution of the sizes of ASes. Our model leads naturally to a method for AS-labelled network topology generation in which the topology grows incrementally, and as nodes are added, new ASes arise, and existing ASes grow and merge.

2 Related Work

We know of no efforts to model the evolution of autonomous system size, although there have been studies of AS sizes and degree distributions, as well as efforts to infer AS-level behavior using records such as BGP tables.

Mahajan et al. [5] develop methods to recognize and characterize BGP misconfigurations in order to determine their effect on connectivity and routing. They recognize misconfigurations by searching for inconsistencies within BGP tables, as well as recording changes in prefix origins. A comprehensive analysis of the evolution of BGP tables comes from the CAIDA group in [2]. Although they provide a detailed taxonomy of the ways in which BGP tables have evolved over time, they do not attempt to model the causes of these trends.

Findings of Faloutsos et al [3] and proposed growth models such as [1] has activated interest in developing more accurate router-level and AS-level topology models and topology generators. For example, the recent work of [9] investigates causes of AS degree distributions and suggest that AS degree may be determined by AS size, but they do not study the causes of AS size distributions.

3 A Simple Growth Model

We motivate our model with observations on the growth of ASes and hosts over time. Using AS number allocations collected from the three Internet registries and estimates of the number of Internet hosts collected from the Internet Domain Survey [4] (we return to justify these choices and present the full details of our methodology in Section 3.2), we plot the growth of ASes and Internet hosts over the last decade in Figure 1. As one might naturally expect, both plots give evidence of exponential growth.

A more interesting finding, which can easily be derived from the two plots in Figure 1 and is also considered in detail in Section 3.2, is that the number of hosts per AS steadily grows over this time period.

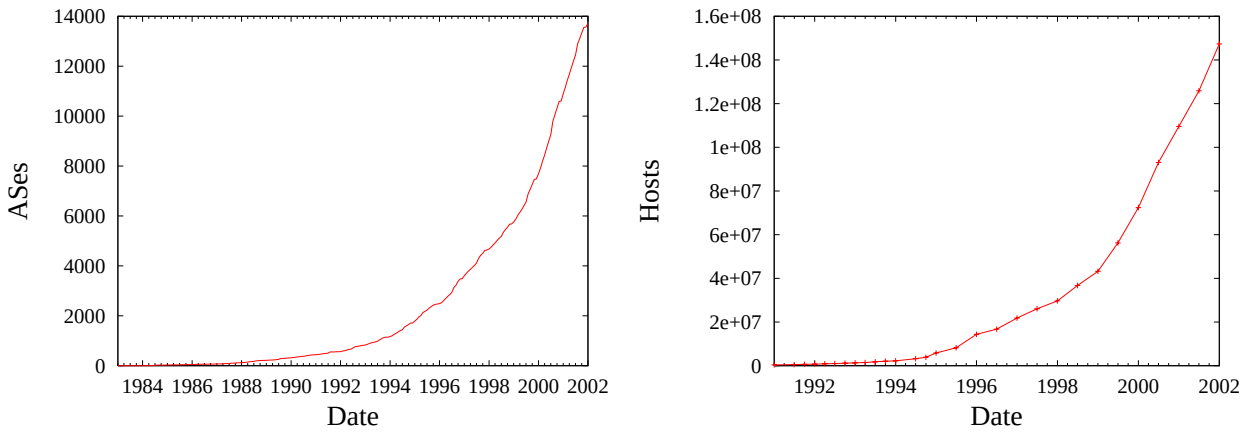

Figure 1: Growth in the number of ASes and Internet hosts.

This growth in the average size of an AS over time gives clear evidence that the size distribution of ASes is not scale-free (in contrast to scale-free distributions produced by other well-known growth models such as [1]). These observations provide starting points for our model.

3.1 A Simple Model and its Analysis

Let $N(t)$ be the total number of ASes ('N' stands for 'number') and $M(t)$ be the total number of hosts ('M' stands for 'mass') in the system. The simplest growth model consistent with the observations above is mathematically described by linear equations

$$\frac{dN}{dt} = qN, \qquad \frac{dM}{dt} = pM + qN. \tag{1}$$

Here q is the rate of creation of new ASes and p is the rate of creation of new nodes. When a new AS is created, the host is given that new label, explaining the qN term in Eq. (1). (For now, we shall assume that there is no merging of ASes; moreover, we assume that links do not affect growth processes and hosts and links never disappear.) Solving for N and M gives

$$N(t) = N(0)\, e^{qt}, \tag{2}$$

$$M(t) = A\, e^{pt} + BN(t), \tag{3}$$

with A, B being simple functions of the initial data $M(0), N(0)$ and the parameters p and q. At the special point $p = q$ the coefficients diverge ($A = B = \infty$), reflecting that the exact solution is actually a linear combination of e^{pt} and $t\, e^{pt}$. Thus the average AS size $\langle s \rangle \equiv M(t)/N(t)$ could in principle exhibit the following asymptotic behaviors:

$$\langle s \rangle \sim \begin{cases} finite & \text{when } p < q, \\ \ln N & \text{when } p = q, \\ N^{(p-q)/q} & \text{when } p > q. \end{cases} \tag{4}$$

We show later (Figure 3) that the average AS size grows over time (and with N), thus the inequality $p > q$ must hold.

Let $N_s(t)$ be the number of ASes with s nodes. This size distribution satisfies the rate equation[1]

$$\frac{dN_s}{dt} = p\left[(s-1)N_{s-1} - sN_s\right] + qN\delta_{s,1}. \tag{5}$$

We already know $N(t) = N(0)\, e^{qt}$. Solving Eqs. (5) recursively and expressing in terms of N rather than t yields

$$N_s = n_s N + \sum_{j=1}^{s} C_{sj} N^{-jp/q}. \tag{6}$$

The coefficients C_{sj} depend on initial conditions while n_s are universal. Asymptotically, only the linear term $n_s N$ matters. To determine this dominant contribution, we insert $N_s(t) = n_s N(t)$ into Eq. (5). We arrive at the recursion relation

$$\left(s + \frac{q}{p}\right) n_s = (s-1) n_{s-1} \tag{7}$$

for $s \geq 2$, while for $s = 1$ we recover $n_1 = q/(q+p)$. A solution to recursion (7) reads

$$n_s = \frac{q}{q+p}\, \frac{\Gamma(s)\, \Gamma\left(2 + \frac{q}{p}\right)}{\Gamma\left(s + 1 + \frac{q}{p}\right)}. \tag{8}$$

Asymptotically, the ratio of gamma functions simplifies to the power law,

$$n_s \sim C\, s^{-\tau}, \tag{9}$$

with $\tau = 1 + q/p$ and $C = \frac{q}{q+p}\, \Gamma\left(2 + \frac{q}{p}\right)$.

3.2 Estimating Growth Rates

In order for us to validate the proposed growth model, we first need to estimate the parameters p and q, the growth rate of the number of hosts and ASes, respectively.

To estimate these rates, we explored a number of alternatives before selecting the methods we deemed most appropriate. For example, BGP tables appear to be a viable alternative for estimating both rates; however, logs only date back to around 1997; moreover, not all IP addresses within a prefix advertised in a BGP prefix are actually in use. The best available method seems to be to use the publicly available

[1] In the large time limit, the random variables $N_s(t)$ become highly localized around corresponding average values.

routing number allocations provided by the ARIN, RIPE, and APNIC registries. Each keeps a public record dating back to the early 1980s of routing number allocations which include, among other details, the routing number and its type (IP or AS), the date on which the number was allocated, the quantity (and in the case of IP allocations, the starting address). It should be noted that RIPE does not publish AS number allocations, though many allocations to that region have been recorded by ARIN.

From these tables we derived the plot of AS growth in Figure 1, and plotted again on logscale in Figure 2(a).[2] Fitting this logscale plot to a line reveals that AS numbers are indeed allocated at an exponentially growing rate. We then estimate q by the slope of the linear regression fit to the curve, or approximately $3.8 \cdot 10^{-4}$.

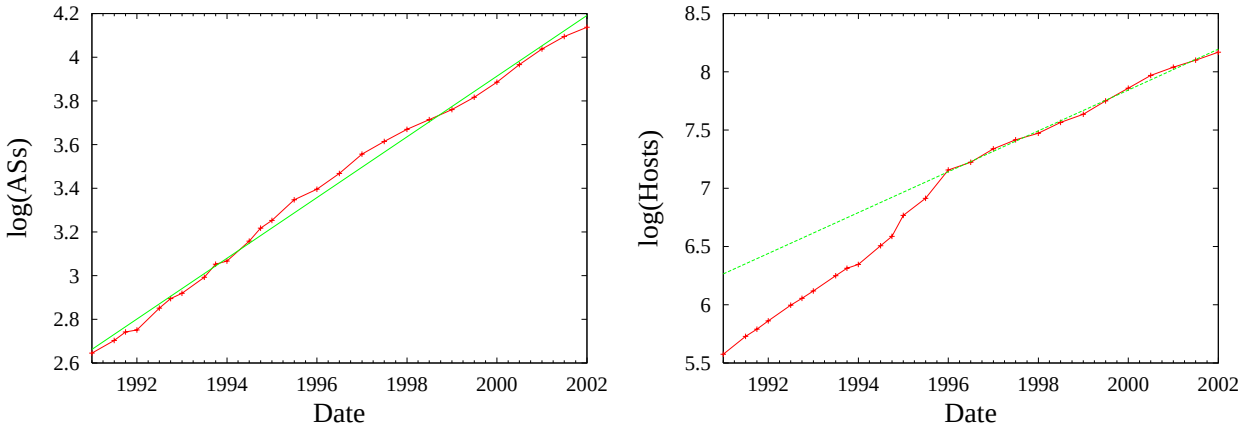

Figure 2: Left a) AS Growth, Right b) Host Growth.

Estimating p, however, is more difficult. The Registries' allocations statistics show that approximately 50% of IP addresses have been allocated and that IP allocations are growing much less than exponentially. As noted below, this trend does not result from the Internet growing less than exponentially, but rather from a growing tendency to better manage allocated IP space.

The Registries provide an excellent record of AS births, but it is infeasible to record IPs in use (and hence, record births of hosts and routers on the Internet). The records of host growth we considered were Telcordia's Netsizer [7] (no longer a publicly available service) and the widely cited Internet Software Consortium's "Internet Domain Survey" project. The host count[3] they develop is based on a reverse DNS process; details can be found at [4].

Using the numbers published by IDS, we plot host growth in logarithmic scale in Figure 2(b). This plot seems to show a change in slope around 1996. Using the more conservative growth rate, *i.e.* the best fit line

[2]Here we assume that an AS typically comes into existence on the Internet shortly after it is allocated, thus the allocations provide a good estimate for q. Also recall that we are primarily interested in the overall *rate* of growth.

[3]We can be certain that Registry IP allocation records do not provide host growth statistics since (using IDS numbers) usage of allocated IP space has increased from less than 1% in 1994 to 8% in 2002.

of the curve following 1996, we find p to be about $4.8 \cdot 10^{-4}$.

We emphasize that while host count may well underestimate the actual number of hosts on the Internet, we are primarily interested in estimating the slope of the curve; our model is unaffected by scaling factors.

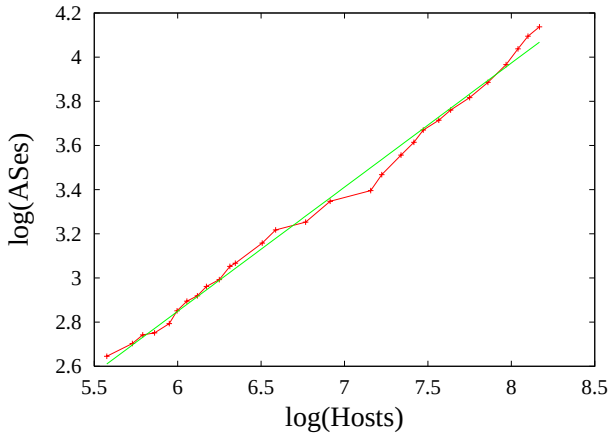

Figure 3: Number of Hosts vs. Number of ASes

Our model also makes a very specific prediction on the relationship between N and M as the system evolves. From Eqn. 4, we expect

$$\langle s \rangle \equiv M(t)/N(t) \sim N(t)^{(p-q)/q},$$

i.e. $M \sim N^{1+(p-q)/q}$. Indeed, we see clear evidence of a power-law fit between M and N when we plot their relationship on log-log scale in Figure 3. The predicted slope is 1.26 and the slope of the linear regression is 0.56, so while the model is in the right ballpark, some additional investigation is warranted.

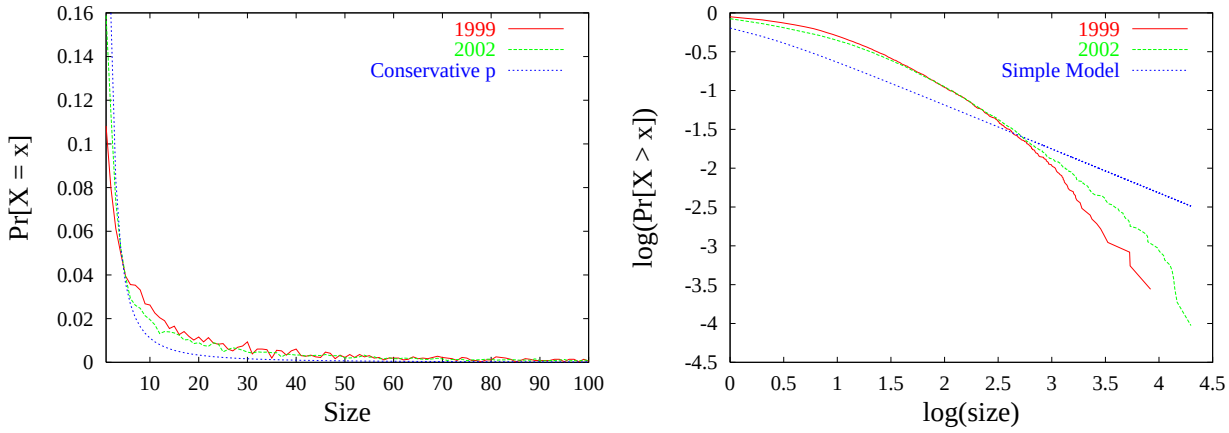

Figure 4: Simple Model Predictions vs. Measurements

We first compare the model's predictions using our estimates for p and q against empirical data from 1999 and 2002 in Figure 4.[4] The pdf is provided to show rough agreement in the body of the distribution, the log-log plot of the ccdf shows the quality of the fit in the tail. We describe our efforts to improve the simple model's predictive power and accuracy next.

4 Modelling AS Mergers

The model described in Section 3.1 is appealing because of its simplicity, but fails to account for a set of prominent events in our datasets — namely, *mergers* between pairs of ASes. In our datasets, we observe these mergers, or coalescence events, in our BGP logs when we witness one AS begin to advertise the set of IP addresses formerly advertised by another AS that then disappears. We provide our methodology for detecting these events in full detail in Section 4.3. Coalescence markedly impacts the manner in which ASes grow, since they enable an AS to grow by a multiplicative factor at a single timestep. In the remainder of this section, we describe how to augment the model to incorporate mergers, analyze the asymptotic behavior of the model and its predictions, and compare the predictions to measurements derived from our data sets.

Now, we shall take into account the increase of the total number of nodes, labels, and the merging between different labels, using the same assumptions as before. The model is now described by linear equations

$$\frac{dN}{dt} = (q-r)N, \qquad \frac{dM}{dt} = pM + qN.$$

The new parameter, r, is the rate of coalescence, i.e. the rate at which two ASes decide to merge. As before, solving for N and M gives differential equations:

$$N(t) = N_0\, e^{(q-r)t} \quad \text{and} \quad M(t) = A\, e^{pt} + BN(t). \tag{10}$$

Following the same analysis of asymptotic behavior as in Section 3.1, and reasoning as before that the average AS size is large and growing, the inequality $p > q - r$ must hold, which implies $\langle s \rangle \sim N^{(p-q+r)/(q-r)}$.

4.1 Implications of the Model

Let $N_s(t)$ be the number of ASes with s nodes. This size distribution satisfies the rate equation

$$\begin{aligned} \frac{dN_s}{dt} &= p\left[(s-1)N_{s-1} - sN_s\right] + qN\delta_{s,1} \\ &+ \frac{rN}{K}\left\{\sum_{i+j=s} K_{ij}N_iN_j - 2N_s\sum_{j=1}^{\infty} K_{sj}N_j\right\}. \end{aligned} \tag{11}$$

[4] Size distributions drawn using Mercator data in 1999, and Skitter data in 2002.

The first term on the right-hand side accounts for growth that proceeds with rate p: When a node is added to an AS with $s-1$ nodes, the number of ASes with s nodes increases by one; similarly when a node is added to an AS with s nodes, the number of ASes with s nodes decreases by one. The next term on the right-hand side of Eq. (11) accounts for nucleation, with rate q, of new ASes (of size one; one can also study more general situations, e.g., sizes of new ASes can be drawn from a distribution). The last term describes coalescence that proceeds with rate r. This term contains a symmetric "kernel" K_{ij}, the rate of merging between ASes with i and j nodes; $K(t) = \sum_{i,j\geq 1} K_{ij} N_i(t) N_j(t)$ is the proper normalization factor.

Our ongoing work focuses on identifying which kernel most accurately reflects actual coalescence behavior. In what follows, we outline the derivation of the asymptotic behavior of the simplest kernel (an exact analysis is provided in the appendix) and briefly motivate a more general class of kernels, which we can also analyze.

4.2 Constant Kernel

Setting $K_{ij} = 1$ transforms Eq. (11) into

$$\begin{aligned} \frac{dN_s}{dt} &= p\left[(s-1)N_{s-1} - sN_s\right] + qN\delta_{s,1} \\ &+ \frac{r}{N}\left\{\sum_{i+j=s} N_i N_j - 2NN_s\right\}. \end{aligned} \tag{12}$$

Equations (12) can be solved recursively. For instance, the number of ASes of the smallest possible size evolves according to $\dot{N}_1 = qN - (p+2r)N_1$. A solution to this equation is a linear combination of two exponents. Asymptotically, the solution simplifies to $N_1(t) = n_1 N(t)$ with $n_1 = q/(p+q+r)$. Similarly, each $N_s(t)$ grows linearly with N. Writing

$$N_s(t) = n_s N(t), \tag{13}$$

we recast Eq. (12) into the recursion relation

$$\begin{aligned} (q-r)n_s &= p\left[(s-1)n_{s-1} - sn_s\right] + q\delta_{s,1} \\ &+ r\sum_{i+j=s} n_i n_j - 2rn_s. \end{aligned} \tag{14}$$

The key is to notice that n_s changes slower than exponentially as $s \to \infty$. This allows to employ the continuum approach, e.g., to replace the difference $sn_s - (s-1)n_{s-1}$ by the derivative $\frac{d}{ds}sn_s$. Using the identity $\sum n_s \equiv 1$ we can also simplify the sum on the right-hand side of Eq. (14) to $2n_s$. Thus for $s \to \infty$, Eq. (14) reduces to the differential equation

$$p\,\frac{d}{ds}\,sn_s = -(q-r)n_s, \tag{15}$$

whose solution has the power-law form

$$n_s \propto s^{-\tau}, \qquad \tau = 1 + \frac{q-r}{p}. \tag{16}$$

Solutions in the special cases $p = q$ and $p = r$ confirm the exponent and additionally give the amplitude; details are provided in Appendix A. We also note that the same analysis which causes the final terms in Eq. (14) to vanish, and reduces that equation to Eq. (15), applies to a much more general class of kernels. We are currently analyzing empirical data in order to determine which kernel best fits our observations.

Note that $\tau < 2$, see Eqn. (16) and recall that $p > q-r$. Interestingly, most of the power laws observed in the Internet are characterized by the exponents exceeding 2. The reason for this is very simple: Distributions P_k, beyond the obvious normalization condition $\sum P_k = 1$, usually satisfy the sum rule $\sum kP_k =$ finite. This sum rule implies that the tail $P_k \propto k^{-\tau}$ must be steeper than k^{-2}. For instance, for the router degree distribution, $\sum kP_k$ is the average degree that is small and does not vary appreciably with time. However, for the AS size distribution, the average size grows indefinitely, implying that $\sum sn_s$ diverges and thence $\tau < 2$.

4.3 Estimating the Rate of Coalescence

Unfortunately, there is no obvious means of tracking AS mergers on the Internet, since we are not aware of any publicly available records providing this information. We therefore resort to making inferences, specifically, by examining aggregated BGP table archives stored by RouteViews [8] at U. Oregon, NLANR and PCH since 1997. Our strategy is to to identify signatures of these merger events from comparisons of sets of daily BGP snapshots This strategy is complicated by the presence of considerable daily churn [2], clouding events of interest with substantial noise.

We argue that aside from churn, there are two reasons for an AS to disappear from daily BGP snapshots.[5]

Coalescence: The IP prefixes formerly advertised by one AS are now advertised by a different AS.

Evaporation: The IP prefixes formerly advertised by one AS simply disappear from the BGP tables.

To infer these two events, we first identify all "suspicious" events on consecutive daily BGP snapshots. We define a suspicious event to be either 1) the occurrence of identical IP prefixes advertised by two different ASes on successive days, or 2) an IP prefix advertised by one AS on one day, followed by a day in which that exact prefix does not appear, moreover the longest matching prefix including the missing prefix is advertised

[5]Using our methods, we cannot detect AS mergers in which the acquiring AS retains use of the acquired AS number as well as its own.

by a different AS. Of course, many of these suspicious events are due to normal BGP churn and its attendant causes. Therefore, the remaining obstacle is to distinguish between an actual merger or disappearance and an instance of BGP churn.[6]

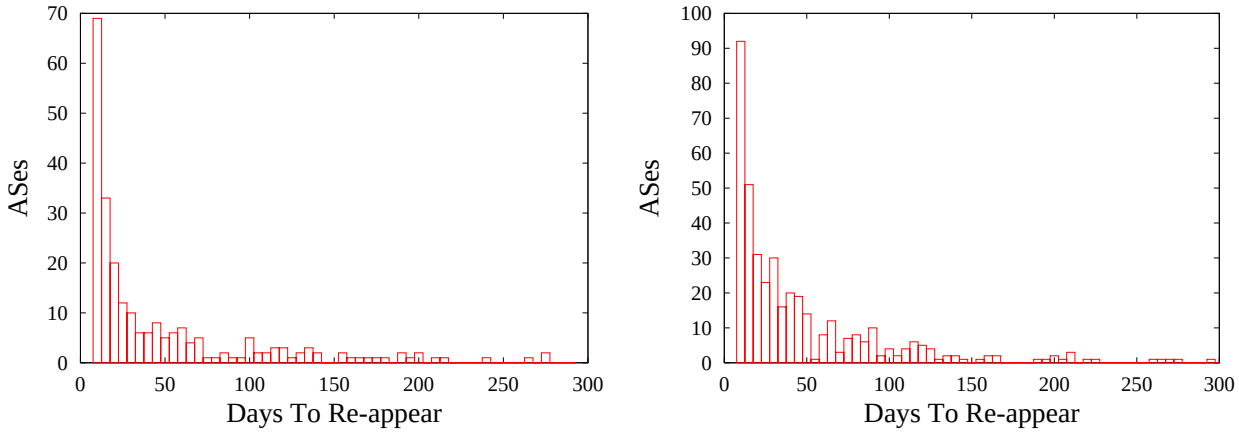

Figure 5: AS Reappearance Time. Left: 2000, Right: 2001.

The distinguishing characteristic we use is to consider the *duration* of time that the AS associated with one or more suspicious events actually disappears from the BGP logs (our concern is that AS numbers are eventually reused). To determine an appropriate cut-off threshold, we measured the duration in days for an AS number to reappear after failing to advertise all of its IP blocks, using RouteViews' BGP tables spanning 01/02/2000 - 11/29/2000, and 02/20/2001 - 02/28/2002. Figure 5 presents histograms (bin width = 5 days) of the time taken for an AS number to re-appear once it has relinquished its IP space. For clarity, we remove the first bin, which clearly corresponds to BGP churn and constitutes the overwhelming majority of disappearances. In total, 89% and 79% of ASes reappeared in the 2000 and 2001 datasets, respectively, and the majority of these returned within a few days of disappearing. For this reason we feel it is reasonable to assume that beyond a cutoff of between three months to a year, the suspicious event is not due to churn. We record a suspicious event as a merger when,

- there is a handover of IP space as discussed above, and
- the AS number losing IP prefixes then disappears, and
- the AS number does not return in the observed interval.

Applying this analysis to the BGP data allows us to measure the quantities needed to estimate r: $L(t)$ is the total number of ASes present in the tables at time t; and $C(t)$ is the total number of ASes that have

[6]Note that it is difficult to distinguish coalescence from evaporation using our methods, since an AS whose IP prefixes evaporate into a larger block of address space is indistinguishable from an AS which coalesces with the AS advertising the surrounding IP block(s).

merged into another AS by time t. Then, using Eqn. (10),

$$\begin{aligned} L(t) &\sim e^{(q-r)t} &\sim (L(t)+C(t))/e^{rt} \\ L(t)/(L(t)+C(t)) &\sim e^{-rt} \end{aligned}$$

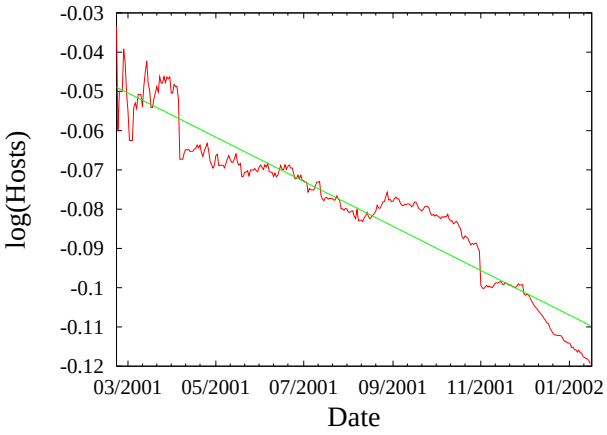

Figure 6: Rate at which ASes Merged in the 2001 Dataset.

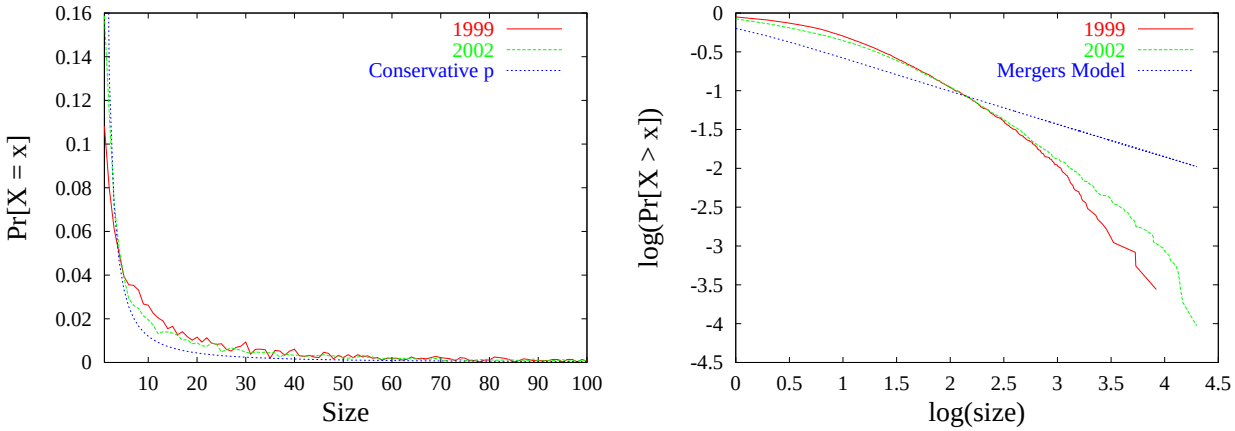

Figure 7: Coalescence Model Predictions vs. Measurements

In Figure 6 we plot $L(t)/(L(t)+C(t)$ on semi-log axes, yielding an estimate $r \approx 1.8 \cdot 10^{-4}$. Figure 7 summarizes the distribution predicted by the coalescence model. As before, only a fit with the body can be seen from the pdf, so a log-log ccdf is provided. Overall, while we find that the complex model is more accurate than the simple in predicting the distribution of small to medium-sized ASes, it still does not give an accurate prediction of large ASes in the tail of the distribution.

5 Conclusions

Understanding the dynamics of AS size distribution is important both for modeling purposes and understanding connectivity. In this paper we have proposed and analyzed two models for the evolution of the AS size distribution.

First, we have provided and analyzed a growth model with rate equations. We have discussed methods for estimating the parameters of this model and shown the size distribution of ASes that it predicts. The model's predictions exhibit size distributions that are in general agreement with empirical data, both in the body and the tail of the distribution. However, discrepancies exist between model and data, particularly in the shape of the tail.

Second, we have suggested that it is important to incorporate the merging of ASes in our models. We show how to do so, and specify the resulting rate equation. The details of this model are highly dependent on assumptions about the manner in which ASes merge, which is captured in the merging kernel – the likelihood that two ASes of given sizes will merge at any timestep. We solve this model for the constant kernel, and show how to estimate the associated parameters. The results point to the need for further analysis of the processes by which ASes merge.

References

[1] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, October 1999.

[2] A. Broido, k. claffy, and E. Nemeth. Internet Expansion, Refinement and Churn. *European Transactions on Telecommunications*, To Appear, 2002.

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *ACM SIGCOMM*, pages 251–62, Cambridge, MA, September 1999.

[4] Internet Domain Survey. At *http://http://www.isc.org/ds/*.

[5] R. Mahajan, D. Wetherall, and T. Anderson. Understanding BGP Misconfiguration. In *To appear in ACM SIGCOMM 2002*, Pittsburgh, PA, August 2002.

[6] P. Meakin. Droplet Deposition Growth and Coalescence. *Rep. Prog. Phys.*, 55:157–240, 1992.

[7] Telcordia's NetSizer. At *http://www.netsizer.com/*.

[8] RouteViews Project at University of Oregon. At *http://archive.routeviews.org/*.

[9] H. Tangmunarunkit, J. Doyle, R. Govindan, S. Jamin, and S. Shenker. Does AS Size Determine Degree in AS Topology? *ACM Computer Communication Review*, 31(5):7, October 2001.

A An Exact Solution For Constant Kernel

Recursions are best analyzed by means of generating functions. Multiplying Eq. (14) by z^s and summing over all $s \geq 1$ we find that the generating function

$$\mathcal{N}(z) = \sum_{s=1}^{\infty} n_s z^s \tag{17}$$

satisfies the Riccati equation

$$pz(1-z)\frac{d\mathcal{N}}{dz} + \mathcal{N}(q + r - r\mathcal{N}) = qz. \tag{18}$$

There is no general technique for obtaining a solution to the Riccati equation. In the present case, however, one can determine the generating function $\mathcal{N}(z)$ by converting Eq. (18) into the hypergeometric equation. The results are combersome so we do not report them here. Instead, we find two special solutions that are valid in a limited parameter range. Recall that Riccati equations can often be solved if one finds a particular solution; then the general solution is easily obtained. The form of Eq. (18) suggests to seek a solution $\mathcal{N}(z) = a + bz$ with yet undetermined constants a, b. The above ansatz works for $p = q$ or $p = r$, and the solutions read

$$\mathcal{N}_1(z) = 1 + \frac{p(1-z)}{r} \qquad \text{when} \quad p = q, \tag{19}$$

$$\mathcal{N}_2(z) = \frac{q}{p} + 1 - z \qquad \text{when} \quad p = r. \tag{20}$$

None of the above "solutions" satisfy the obvious boundary condition $\mathcal{N}(0) = 0$; the second solution also does not satisfy the boundary condition $\mathcal{N}(1) = 1$. However, we can find the general solution, and then choose the true solution. We now implement this program in detail for the case of $p = q$. According to standard procedure, we seek the general solution $\mathcal{N}(z) = \mathcal{N}_1(z) + y(z)$ and find that $y(z)$ satisfies a Bernoulli equation. In the present case, the Bernoulli equation reads

$$pz(1-z)\frac{dy}{dz} = ry^2 + (p + r - 2pz)\, y. \tag{21}$$

The general solution of the Bernoulli equation (21) is

$$y(z) = \frac{z^{1+\frac{r}{p}}(1-z)^{1-\frac{r}{p}}}{C_0 - \frac{r}{p}\int_0^z dx\, x^{\frac{r}{p}}(1-x)^{-\frac{r}{p}}}\,. \tag{22}$$

We must choose $C_0 = 0$ to ensure that $\mathcal{N}(0) = 0$. Thus the solution of the Riccati equation is

$$\mathcal{N}(z) = 1 + \frac{p(1-z)}{r} - \frac{p}{r}\,\frac{z^{1+\frac{r}{p}}(1-z)^{1-\frac{r}{p}}}{\int_0^z dx\, x^{\frac{r}{p}}(1-x)^{-\frac{r}{p}}}\,. \tag{23}$$

Expanding $\mathcal{N}(z)$ one can recover n_s for any s. There appears to be no closed-form elementary expression. Exact formulas for n_s become very unwieldy as s increases, yet one can establish a simple asymptotic formula. indeed, the asymptotics of the size distribution can be read off from the behavior of the generating function in the $z \to 1$ limit. From Eq. (23), we have

$$\mathcal{N}(z) = 1 - \frac{p/r}{B(1+r/p, 1-r/p)}\,(1-z)^{1-\frac{r}{p}} + \mathcal{O}(1-z),$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Euler's beta function. Such singular behavior shows that the distribution n_s has a power law tail. Specifically,

$$n_s \sim C\, s^{-\tau} \tag{24}$$

with $1 < \tau < 2$ would imply that the generating function (17) behaves as

$$\mathcal{N}(z) = 1 + C\Gamma(1-\tau)\,(1-z)^{\tau-1} + \mathcal{O}(1-z) \tag{25}$$

when $z \to 1$. By matching terms of order of $(1-z)^{\tau-1}$ in the above two expansions of $\mathcal{N}(z)$ we obtain the exponent, $\tau = 2 - r/p$ in agreement with the general prediction of Eq. (16), and additionally the amplitude

$$C = \frac{1}{\Gamma(3-\tau)}\,\frac{(\tau-1)\,\sin\pi(\tau-1)}{\pi\,(2-\tau)}, \qquad \tau = 2 - \frac{r}{p}.$$

Consider now the complementary case $p = r$. We write $\tau = q/p$ since, as we shall see in a moment, the exponent τ is indeed equal to q/p in agreement with the general prediction of Eq. (16). Employing the same approach as earlier we arrive at

$$\mathcal{N}(z) = \tau + 1 - z - \frac{z^{1+\tau}(1-z)^{1-\tau}}{\int_0^z dx\, x^{\tau}(1-x)^{-\tau}}\,. \tag{26}$$

Since $q > r = p$, the integral in the denominator on the right-hand side of Eq. (26) diverges in the $z \to 1$

limit. It proves useful to re-write the integral as

$$\int_0^z dx\,(1-x)^{-\tau} + \int_0^1 dx\,\frac{x^\tau - 1}{(1-x)^\tau} - \int_z^1 dx\,\frac{x^\tau - 1}{(1-x)^\tau}.$$

The first (diverging) integral equals

$$\frac{(1-z)^{1-\tau} - 1}{\tau - 1}.$$

The second integral equals

$$\frac{\pi\,\tau}{\sin(\pi\,\tau)} + \frac{1}{\tau - 1},$$

and the third integral vanishes as $(1-z)^{2-\tau}$. Now we expand $\mathcal{N}(z)$ in the $z \to 1$ limit, and compare the outcome with (25). The first non-trivial term is indeed $(1-z)^{\tau-1}$ thus confirming the value of the exponent. Additionally, we obtain the amplitude

$$C = \frac{1}{\Gamma(2-\tau)}\,\frac{\pi\,\tau\,(\tau-1)^3}{\sin\pi(\tau-1)}, \qquad \tau = \frac{q}{p}.$$

Thus for $p = q$ and $p = r$ we have computed the generating function[7]. These solutions provide rigorous confirmation of the general expression for the exponent, and also give the precise asymptotics.

[7] Solution comes with help the following reference, C.M. Bender and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers.* McGraw-Hill Book Co., Singapore, 1984